Original Article

Healthc Inform Res. 2011 March;17(1):24-28.
doi: 10.4258/hir.2011.17.1.24
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Evaluation of Co-occurring Terms in Clinical Documents Using Latent Semantic Indexing

Choonghyun Han, MS[1], Sooyoung Yoo, PhD[2], Jinwook Choi, MD, PhD[3]

[1]Interdisciplinary Program of Bioengineering, College of Engineering, Seoul National University, Seoul; [2]Seoul National University Bundang Hospital, Seongnam; [3]Department of Biomedical Engineering, College of Medicine, Seoul National University, Seoul, Korea

Objectives: Measurement of similarities between documents is typically influenced by the sparseness of the term-document matrix employed. Latent semantic indexing (LSI) may improve the results of this type of analysis. **Methods:** In this study, LSI was utilized in an attempt to reduce the term vector space of clinical documents and newspaper editorials. **Results:** After applying LSI, document similarities were revealed more clearly in clinical documents than editorials. Clinical documents which can be characterized with co-occurring medical terms, various expressions for the same concepts, abbreviations, and typographical errors showed increased improvement with regards to a correlation between co-occurring terms and document similarities. **Conclusions:** Our results showed that LSI can be used effectively to measure similarities in clinical documents. In addition, correlation between the co-occurrence of terms and similarities realized in this study is an important positive feature associated with LSI.

Keywords: Information Storage and Retrieval, Cluster Analysis, Documentation

I. Introduction

The spread of electronic medical records increases utilization of clinical experience, knowledge and information contained in clinical documents. To search, collect and retrieve relevant information according to the user's need is one of the important issues in current medical informatics.

Various kinds of information technologies have been applied to medical domain to improve retrieval's efficacy. Measurement of similarity using vector space model is one of the widely used basic methods. Even though the vector space model is both efficient and effective, sparseness is a problem deteriorating its performance of the retrieval. When composing a document, people use various expressions, and the use of synonyms or acronyms is very common.

Conventional vector space model regards each term as a separate dimension axis. As we use many kinds of different expressions on the same thing, there will be a lot of various axes regarding one subject. Because of the variety of expressions, Landauer [1] pointed out that 99% of cells in term-document matrix were vacant. In addition, we assume that clinical documents will be generated using clinical jargons which are relatively small number compared with everyday language terms.

In this study, we review a method in order to measure the similarity among documents. In order to see the nature of similarity among various documents, we applied latent semantic indexing (LSI) to newspaper editorials and clinical documents. Finally, we analyzed major variables affecting the similarity.

Received for review: July 20, 2010
Accepted for publication: March 5, 2011

Corresponding Author
Jinwook Choi, MD, PhD
Department of Biomedical Engineering, College of Medicine, Seoul National University, 28 Yeongeon-dong, Jongno-gu, Seoul 110-799, Korea. Tel: +82-2-2072-3421, Fax: +82-2-745-7870, E-mail: jinchoi@snu.ac.kr

Deerwester et al. [2] proposed LSI using singular value decomposition (SVD), and compared the performance with that of SMART system. Through analyzing previous latent semantic analysis (LSA) and information retrieval researches of her and other scholars, Dumais [3] found the critical factors affecting the performance of LSA. They are the number of dimensions to reduce to and the diversity, and the size of the collection, and the number of singular values extracted (Figure 1).

Hofmann suggested probabilistic LSI, which accepted basic idea of LSI, and applied the probabilistic model and expectation maximization algorithm instead of SVD [4,5]. Adopting LSI and pLSI, Blei et al. [6] developed latent dirichlet allocation (LDA), as a modified probabilistic topic model. Many kinds of variations were studied since then [7-9].

There are three usages related to the LSI. One approach is to regard LSI as a soft clustering method [10]. Another is to emphasize the role of dimension reduction. And the third is to focus on the solution to the matter of sparseness. LSI is said to be a method to solve the sparseness problem. Instead of hyperdimensional space composed of terms' axis, terms and documents are located on relatively low dimensional space created by LSI.

II. Methods

1. Data Collection

15,618 deidentified discharge summaries of Seoul National University Hospital of year 2003 were collected. They include information such as patients' gender, age, admission date, discharge date, ward, department, doctors in charge, chief complaint, onset of problems, patient status when admitted, diagnosis, problem list, test results, discharge type, prescriptions, appointment, history, results of physical examination, progress, things to care about after discharge, operation record, and plan for future. Some of those columns are filled with free text description, and others contain predefined coded data.

At each trial, we randomly selected 1,000 documents from the collection. At one trial, we collected data from all kinds of available columns, and another trial, we collected data from columns composed of only free text description. All discharge summaries used in this study consist of the mixture of Korean and English words.

As the second document collection we collected newspaper articles which we assume having variety of expressions and terms. We collected 1,000 editorials from three Korean daily newspapers published in either 2008 or 2009. All documents were written in Korean, some of which include number, alphabet characters or words, or traditional Chinese letters.

We classified document sets into three groups; ED, CF, and CS. 'ED' means the collections of newspaper editorials. CF means clinical documents with full columns. CS means clinical documents with selected columns. While CF contains all sections in the discharge summary, but CS has only limited sections such as chief complaint, patient state status of patients when admitted, diagnosis, problem list, test results, discharge type, history, results of physical examination, progress, things to care about after discharge, operation record, and plan for future.

2. Preprocessing

Korean is classified as an agglutinative language. Most words are formed by joining morphemes together [11]. Morphemes in Korean are usually divided into lexical morpheme and grammatical morpheme. One has meaning, but the other just plays grammatical role in a sentence. While prepositions in English are written as separate words in a sentence, grammatical morphemes in Korean are attached to lexical morphemes in the token.

In this study, only lexical morphemes are regarded as meaningful terms, and grammatical morphemes are discarded, as stop words are ignored in many information retrieval tasks. Lexical morphemes in Korean take similar parts to those played by word stems in English at the task of information retrieval using bag-of-words method.

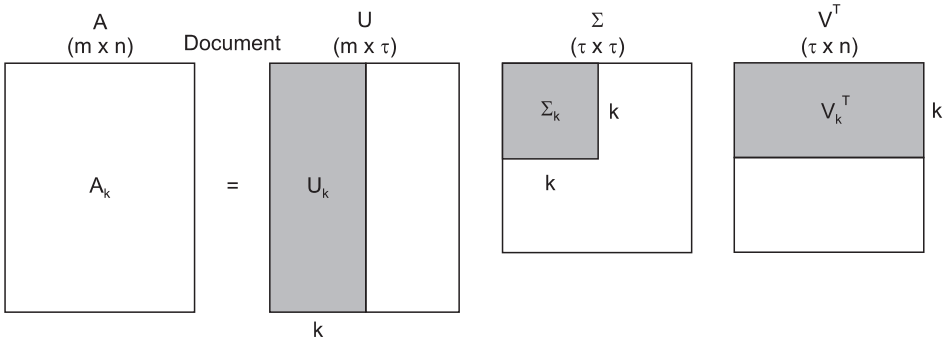

Figure 1. Diagram showing a conceptual description of singular value decomposition.

Table 1. Number of terms and proportions of zero-filled cells

	No. of unique terms	No. of documents	Zero-filled cells (%)
ED	17,077	1,000	99.0
CF	14,610	1,000	98.9
CS	13,809	1,000	99.0

ED: editorials, CF: clinical documents with full columns, CS: clinical documents with selected columns.

Table 2. Cosine values which are calculated with term frequency (TF) and inverse document frequency (IDF) increased after latent semantic indexing (LSI)

	Mean	SD
Cosine value before LSI (with TF-IDF)		
ED	0.026	0.035
CF	0.034	0.031
CS	0.032	0.030
Cosine value after LSI		
ED	0.195	0.121
CF	0.353	0.167
CS	0.256	0.128
Increase of cosine value after LSI		
ED	0.169	0.099
CF	0.319	0.155
CS	0.224	0.110

SD: standard deviation, ED: editorials, CF: clinical documents with full columns, CS: clinical documents with selected columns.

In most cases, clinical documents contain various kinds of abbreviations. Most acronyms included in the clinical documents were accepted without any change, but very common terms with just one letter were replaced with longer forms. For example, "A/N/V/D/C (+/-/-/+/-)" were replaced with "Anorexia/Nausea/Vomiting/Diarrhea/Constipation (+/-/-/+/-)". Typos in the clinical documents were not corrected for this study.

3. Singular Value Decomposition

At first, term-document matrices were built. Cells of the matrices were filled with term frequencies of each lexical morpheme. MATLab 7.01 (MathWorks, Natick, MA, USA) was used for the singular value decomposition of term-document matrices. All three kinds of collections were reduced to 100 dimensions. In this study, we chose 100 as the dimension size, considering both Deerwester et al. [2]'s research records. R 2.10.0 was used for the statistical analysis and Microsoft SQL Server (Microsoft Corporation, Redmond, WA, USA) 2008 was used for the flexible management of data.

Table 1 shows the general information of collections' composition. Landauer's assertion that about 99% of cells of term-document matrix are filled with 0 can be proved in these collections.

4. Co-occurrence

We set three kinds of operational definitions for the co-occurrence of terms between documents to measure their influence on documents' similarity. The first of them is the number of shared terms. Number of shared terms between documents means how many kinds of unique terms are shared between two documents. The second variable is the averaged shared term frequency. Shared term frequencies are summed, and divided by the number of unique terms shared by both documents. The third variable is the averaged unshared term frequency between documents.

With these variables, this study tries to check the relation between the co-occurrence of terms and documents' similarity empirically with Pearson's correlation.

III. Results

1. Document Similarity Measurement

Similarity between two documents was measured using the vector space model. It was calculated before and after LSI as described in [2]. All documents in a collection are combined as pairs, and their similarity is calculated according to the method above. From each collection, the similarities of 499,500 document pairs were measured.

Table 2 shows that the increase of cosine values between documents after LSI in clinical documents is more remarkable than editorials. Average similarity between clinical documents with full column (CF) is the highest among the three collections.

2. The Characteristics of Co-occurrence among Different Collections

Table 3 shows the results of measuring term co-occurrence among various collections. Number of terms shared by two documents showed large difference. Clinical documents share more terms than editorials. The average number of shared terms are 29.37 for clinical documents containing full columns, 20.52 for clinical documents having simple columns, and 16.57 for editorials.

Figure 2 shows the distribution of co-occurring terms among collections. As shown in Table 2, the average number

doi: 10.4258/hir.2011.17.1.24

Table 3. Number of co-occurring terms. Terms from editorials (ED), clinical documents with full columns (CF), and clinical documents with selected columns (CS)

		No. of co-occurring terms	Average shared TF	Average unshared TF
ED	Mean	16.57	1.18	1.34
	SD	6.27	0.15	0.08
	Min.	0.00	0.00	1.05
	Max.	180.00	3.07	1.77
CF	Mean	29.37	1.33	1.37
	SD	16.99	0.27	0.12
	Min.	2.00	1.02	1.00
	Max.	166.00	4.00	2.17
CS	Mean	20.52	1.14	1.35
	SD	16.51	0.18	0.13
	Min.	0.00	0.00	1.00
	Max.	180.00	5.00	2.77

TF: term frequency, SD: standard deviation, Min: minimum, Max: maximum.

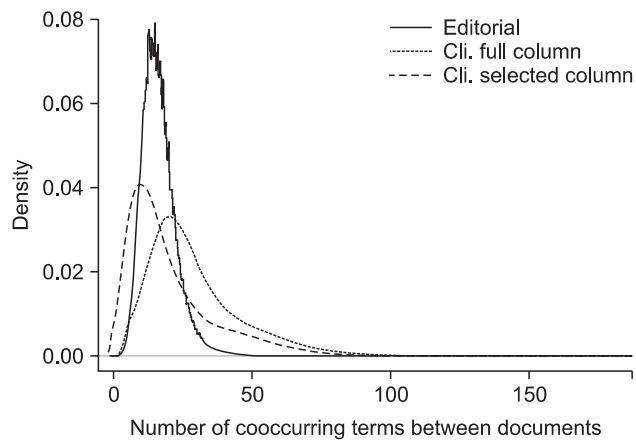

Figure. 2. Distributions of unique number of shared terms in editorials and clinical (Cli.) documents.

of co-occurring terms are slightly higher in clinical documents than editorials. That indicates that clinical documents are more generated based on a bit confined jargons than editorials. The nature of using domain specific jargons can be used for the future study in the field of clustering or extracting topics from document collections.

3. Evaluation of Co-occurring Term Influence

To check the correlation between the co-occurrence and the similarity between documents, we measured Pearson's correlation coefficients between each operational variable defined above and cosine similarities. According to the Pearson's correlation analysis, relations between each variable and cosine similarities show different aspect by collections (Table 4).

Table 4. Pearson's correlation between number of co-occurring terms and document similarity

		Pearson's correlation	*p*-value
ED	Unique term	0.628	<0.001
	Avg. shared TF	0.559	<0.001
	Avg. unshared TF	-0.514	<0.001
CF	Unique term	0.443	<0.001
	Avg. shared TF	0.662	<0.001
	Avg. unshared TF	-0.417	<0.001
CS	Unique term	0.624	<0.001
	Avg. shared TF	0.364	<0.001
	Avg. unshared TF	-0.197	<0.001

ED: editorials, CF: clinical documents with full columns, CS: clinical documents with selected columns.

In ED, the number of co-occurring terms has fairly good correlation with document similarity. In contrary, in CF the correlation between term co-occurrence and similarity is low. However, in CS the correlation shows fairly high level as that of ED. So we can say that the effect of co-occurrence on document similarity can be regarded similar between editorial collections and clinical collections of selected columns. The reason for showing low correlation of clinical collections of full columns is reviewed in discussion section.

IV. Discussion

In this study, we evaluated the importance of co-occurrence on the document similarity. In order to check whether co-occurrence of terms plays a key role on document similarities, we performed the experiment using three types of collections. Through the experiments, we found out that co-occurrence of terms explain large portion of the similarity among documents.

The results reveal the domain specific document characteristics. Similarities between documents were generally higher in clinical documents than editorials, which can be interpreted that in clinical field the domain specific jargons usage is more popular than newspaper arena.

The low correlation of co-occurrence and similarity in CF implies a significant meaning. CF collection has all kinds of columns (subsections) of discharge summary compared with CS which has only selected subsections. Subsections such as department, ward, and admission date can have many co-occurring terms over all clinical documents. Those non-

specific generally common terms can affect as noisy terms in the collection. In contrary CS does not have those sections, which leads to better result in the correlation test.

The dimension size of 100 was set after a series of dimension setting experiments. However, there was no general rule for the determination of proper dimension size. Although Dumais [3] did not come up with general criterion for reasonable dimension size for LSI, but according to her experience, the dimension size should be large enough to yield fairly good performance. In her research, she found out that the information retrieval performance in medical collection was peak at the dimension size of 90.

LSI is theoretically based on the co-occurrence of terms between documents. However, co-occurrence alone cannot explain the effectiveness of LSI in most cases. According to Landauer's study [1], the correlation between LSA-measured word pair similarities and the number of times they appeared in the same passage was only a little higher than that with the number of times they appeared separately in different passages. He asserted that the number or proportion of literal words shared between two passage is not the determinant of their similarity in LSA [1].

Similarity between documents measured after LSI was higher than that with term frequency-inverse document frequency matrix. Not only the average but also increase of cosine values was higher in clinical documents. Mathematically, LSI seems to be effective in exaggerating the similarity between documents. But, the similarity was not evaluated through the comparison with the relevant documents set. To check whether LSI is really better method for the measurement of similarity between clinical documents, further study with gold standard is needed.

The need of finding the best method for clustering clinical documents triggers this research. As we expected there would be lots of co-occurring jargons in medical field compared with other fields, we performed an experiment showing the effect of term co-occurrence on document similarity. Through the experiment, we found out that high frequencies of term co-occurrence in clinical documents were highly associated with similarity among documents. The effect was more significant in clinical collections than editorial collections. However, in clinical collections there are huge number of non-specific co-occurring terms which can be regarded as noise terms, we suggest that for the refined further study, the preprocessing should be considered seriously.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2009-0075089).

References

1. Landauer TK. LSA as a theory of meaning. In: Landauer TK. McNamara DS, Dennis S, Kintsch W, eds. Handbook of latent semantic analysis. Mahwah (NJ): Lawrence Erlbaum Associates; 2007: p3-35.
2. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci 1990; 41: 391-407.
3. Dumais ST. LSA and information retrieval: back to basics. In: McNamara DS, Dennis S, Kintsch W, eds. Handbook of latent semantic analysis. Mahwah (NJ): Lawrence Erlbaum Associates; 2007: p293-321.
4. Hofmann T. Probabilistic latent semantic indexing. In: 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; 1999 Aug 15-19; Berkeley, CA. New York: Association for Computing Machinery; 1999. p50-57.
5. Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 2001; 42: 177-196.
6. Blei D, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003: 3; 993-1022.
7. Steyvers M, Smyth P, Rosen-Zvi M, Griffiths TL. Probabilistic author-topic models for information discovery. In: Knowledge Discovery and Data Mining 2004; 2004 August 22-25; Seattle, WA. p306-315.
8. Wang X, McCallum A. Topics over time: a non-Markov continuous-time model of topical trends. In: Knowledge Discovery and Data Mining 2006; 2006 August 20-23; Philadelphia, PA. New York: Association for Computing Machinery; 2006.
9. Blei DM, Lafferty JD. Dynamic topic models. In: The 23rd International Conference of Machine Learning; 2006 June 25-29; Pittsburgh, PA.
10. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. 1st ed. New York: Cambridge University Press; 2008. p378-384.
11. Wikipedia. Agglutinative language. Wikipedia; [cited at 2010 Feb 19]. Available from: http://en.wikipedia.org/wiki/Agglutinative_language.

 doi: 10.4258/hir.2011.17.1.24